COGNITIVE STUDIES | ÉTUDES COGNITIVES, 23
Warsaw 2023

Article No.: 3027
DOI: 10.11649/cs.3027

Citation: Bouressace, H. (2023). Computational analysis of printed Arabic text database for natural language processing. *Cognitive Studies | Études cognitives*, *2023*(23), Article 3027. https://doi.org/10.11649/cs.3027

Hassina Bouressace

Department of Computer Science, 8 Mai 1945 – Guelma University, Guelma, Algeria
bouressace.hassina@univ-guelma.dz

Computational Analysis of Printed Arabic Text Database for Natural Language Processing

Abstract

A frequency dictionary of printed Arabic text is essential for natural language processing. It includes 1,251 XML files of Arabic documents collected from ten newspapers and magazines from different countries and created as the PATD database. A total of 2,344 articles were created with various structures: open vocabulary, multi-font, multi-size, and multi-style text. From these articles, 1,102,078 tokens, 19,926 sentences, and 1,000,000 words were extracted. This dictionary provides detailed information for each word, including English equivalents, usage statistics, usage distribution, and the most widely used terms. A thematic vocabulary list of the top words on various topics is also provided. This frequency dictionary is a useful resource of modern Arabic vocabulary for various specialists, students, and learners. The frequency dictionary is freely available to interested researchers on the webpage.

Keywords: Arabic language; vocabulary; Arabic documents; frequency dictionary; Arabic printed text database

1 Introduction

Regardless of how the world has evolved over the years and despite the existence of artificial intelligence such as deep learning algorithms, words remain the most valuable tool for transmitting information among people. All print media uses words to spread news and information via articles and sentences. The vocabulary of these sentences is based on specified frequency information that contains a combination of appropriate words, style, and content. Thus, proper vocabulary is an essential part of various fields. One of these fields is natural language processing (NLP). Many models of this technique require massive amounts of labeled data for the natural language processing algorithm to train on, identify relevant connections, and assemble. This kind of big data set is one of the main hurdles to natural language processing. In addition, the process of the NLP technique becomes more effective and authentic if the training data used is accurate. Various algorithms and systems have been created and developed and much progress has been made in natural language processing; therefore, systematic testing frequency information needs to be established by extracting the most frequent words from different documents and creating a thematic vocabulary list of the top words from a variety of key topics.

Although the training data for natural language processing is widely available and its techniques are highly developed in many fields, there are few frequency dictionaries of Arabic today. In spite of this, a number of studies have been carried out in the area of dictionary frequency in Arabic newspapers. Because document and text recognition are related in the present as projects for researchers to enhance the comprehension of the text starting from images, we chose a printed

Arabic text database that contains captured images of newspapers and its XML files, which consists of the text of these images. The study highlights the importance of accurate word selection in Arabic newspapers and magazines. Furthermore, the frequency of words and their significance change over time, which encourages us to conduct new studies and create a new dictionary with up-to-date information in order to extend the training data to enhance the quality of assessment in Arabic language processing. Several previous studies (Abdelali et al., 2005; Adham et al., 2009; Alderson, 2007; Dornyei, 2007) have proven that basic words can be specified by extracting the most frequent words in the texts and then ordering them from the most frequent words to the least frequent, listing them from top to bottom.

The rest of the paper is organized as follows: In Section 2, we survey the Arabic frequency dictionaries available in the literature. In Section 3, we present some Arabic script properties and a variety of types of Arabic printed script, and we give the details of the Arabic database used in this study. In Sections 4 and 4.1, we provide a statistical analysis of all aspects relating to the dictionary. Lastly, we present our key conclusions in Section 5.

2 Literature Review

A considerable number of studies have focused on the frequency of information in Arabic media texts (newspapers and magazines). In this section, we shall discuss some of the recent studies. The purpose of all the previous studies was to help researchers in various fields, such as grammar, language variation, linguistics, and language acquisition.

Al-Sulaiti and Atwell (2006) noted that only one Arabic corpus is publicly available. The researchers mentioned that creating a frequency dictionary using digital information has boosted frequency analysis of Arabic newspapers. Abuleil and Evans (2002) presented pre-computer era word frequency studies on Arabic newspapers by developing a new parser system to read Arabic newspaper articles by generating several sets of rules and techniques and adding new words and features to the dictionary every time they analyze the sentence using 100 articles (80,444 words) from the Al-Raya newspaper. Another Arabic corpus was created by Ahmed Abdelali (2003), which was collected from ten Arabic newspapers in different aspects (spelling, foreign words, and word usage). This study intends to extract the differences in word spelling, imported words, and word usage. The study shows that differences occur between Arabic newspapers in different countries. However, this corpus is limited to noun classification, neglecting other parts of grammar, such as verbs and adverbs. Abdul (Abdul Razak, 2011) presented a comparative analysis of language style using 30 articles from seven Arabic newspapers. This study provides the differences and similarities among the words by applying the likelihood ratio test. It shows that there were small differences in the words' properties, such as word spelling, loan words, and verb transitivity. They discovered that not all sections of newspapers have the same guidelines for writing reports and newspaper articles. For example, the sports section was more specific than the world affairs section. Another Arabic dictionary was created by Buckwalter and Parkinson (Buckwalter & Parkinson, 2011) in written and spoken modes to extract the 5,000 most frequent words, plus the most widely spoken Arabic dialects using a 30 million-word corpus. A thematic vocabulary list containing 30 key topics is available in this dictionary, along with a frequency index where each word has several features. According to Masrai and Milton's Arabic lemmatized frequency lists (Masrai & Milton, 2016), the most common 9,000 lemmatized words provide almost 95% coverage, and the most frequent 14,000 words provide nearly 98% coverage, which they created from a large web-based corpus. The lemma used in their study is more relevant to European languages than Arabic, showing that the relationship between word frequency and coverage in Arabic is comparable, to a certain degree, to English and Greek but not to French; however, the lemma used is very specific to a small cluster. Goweder and De Roeck (2001) generated an 18.5 million-word corpus from Al-Hayat newspaper text, with articles tagged. They compared their findings to English-language descriptions of experiments that had previously only been managed on small, manually collected datasets. They conclude that there are much fewer data in English for the same amount of data in Arabic. Many researchers have used frequency dictionary data to improve document classification, one of which is the Alhaj et al. study (Alhaj et al., 2018), which focused on the influence of word frequency in Arabic document classification and its effects on the representation of characteristics, specifically the bag of words (Bow) (Uysal & Gunal, 2012) and term frequency (TF-IDF) (Ayadi et al., 2016), using three classification techniques, namely Naive Bayes (NB) (El Kourdi et al., 2004), k-nearest neighbor (KNN) (Alshammari, 2018; Syiam et al., 2006), and Support Vector Machine (SVM) (Mesleh, 2007). The results demonstrate that the SVM classifier was upgraded to KNN and NB classifiers using the TF-IDF representation approach and that the NB classifier surpassed the KNN and SVM classifiers when using the representation approach in Bow. Other studies (Duwairi et al., 2009; Goweder & De Roeck, 2001) demonstrate that reducing the high dimensionality of feature space on document classification may improve the efficiency of classifier algorithms, such as by removing a significant amount of non-informative data, which will save valuable processing time and space while also improving accuracy.

3 Data Size/Features

3.1 Features of the Arabic Language

Arabic has unique features compared to other languages, which affects the efficiency of natural language processing application. These features make automatic processing difficult, and they arise from the unique nature and the varied meaning of Arabic words according to their position in the sentence. One of these characteristics is that a word with the same letters can have more than one meaning and can have up to five or six meanings (multi-context), where the word's position in the sentence is the identifying tool of the correct context due to the absence of diacritics. In Arabic, diacritics have the same roles as vowels. However, the Arabic community has become accustomed to understanding the context of a word based on its presence in the sentence without diacritics. This complexity indicates that one incorrect context can affect the entire sentence, which makes it a challenging task that requires a large amount of data to achieve precise results. In addition, Arabic words are very often not separated by spaces. This divergence occurs more frequently in Arabic than in English. As an example, the phrase "by his hand" contains three words in English, whereas in Arabic it is just one word.

3.2 Database

In this statistical study, we used the PATD database (Bouressace & Csirik, 2019). The PATD contains two datasets; each dataset was compiled from ten newspapers and included 200 newspaper pages. The document images were divided into article blocks that may contain one article or several articles. Table 1 shows the percentage of elements that were used in this study.

Based on the nature of the XML file, we have two sets. The first covers all the tokens that can exist in the file, whether they are text or not. The second set covers only the words, leaving out all symbolic characters and numbers. Also, the XML file includes two languages: Arabic and English. Therefore, the English words are ignored. In this study, we will use 1,251 XML files for extracting the frequency wordlist, with each XML file containing one or many articles. The average word count is 425 words in each file, including headlines, sub-headings, authors, etc.

4 Alphabetical Index

This section presents the findings of the analysis of this study. Using the sketch engine tool (Kilgarriff et al., 2004, 2014), we were able to extract the most important statistics from the whole corpus, which provided the necessary information and data for each level. These statistics can be

Table 1. Sub-corpus sizes.

Name	Tokens	Words	%
Akhersaa	63,257	57,397	5.7
Alayam	128,935	116,992	11.7
Alhadaf	135,356	122,818	12.3
Alnahar	95,866	86,986	8.7
Alriad	112,520	102,097	10.2
Alsharek	95,863	86,983	8.7
Alsharek al-Awsat	107,159	97,233	9.7
Alshorouk almisri	88,379	80,192	8.0
Alshourouk	220,239	199,839	20.0
Aswaq qatar	54,504	49,455	4.9

divided into two major classes. The frequency list of words in this corpus is presented in the first class. Along with the rank frequency (R), English equivalent, document frequency (DOCF), relative document frequency (RDOCF), average reduced frequency (ARF), and average logarithmic distance (ALDF), we also provide the most frequent prepositions, conjunctions, verbs, and nouns that were used in this corpus. Another statistic covered the building blocks of corpus text that contain the basic blocks of the sentence parts, including nouns, prepositions, subordinating conjunctions, verbs, adjectives, punctuation, pronouns, coordinating conjunctions, cardinal numbers, particles, determiners, adverbs, and interjections. In the second class, we present the N-gram and keywords analysis by extracting the most contiguous sequence used at each gram level, providing the necessary information as in the first part, such as RDOCF and ALDF.

4.1 Frequency Wordlist

The figure below (1) shows the distribution of words by frequency, which contains 51,847 different words listed from highest frequency to lowest frequency without counting the XML element words, dividing the word frequency into many sets.

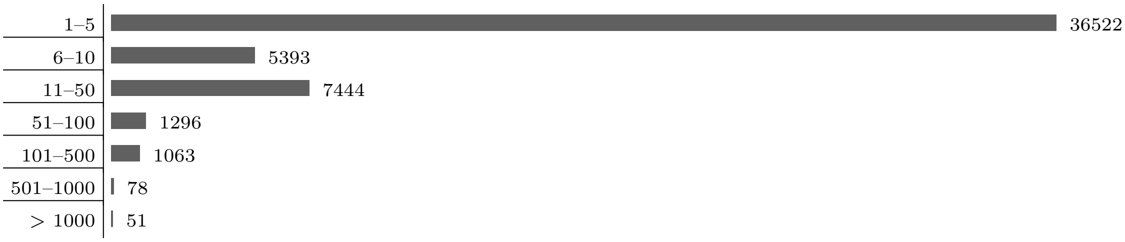

Figure 1. Number of words by frequency from top to bottom of the list.

From this graph (Figure 1), we can observe that the differences among the sets are significant. The words that have fewer than or equal to five occurrences make up more than 70% of the corpus, while the words that have more than 1,000 occurrences cover 51 words. Based on this data, we conclude that there is a specific set of words that are used frequently in the newspaper, while low-frequency words are generally used to highlight the indication in special cases. This data can be divided into two parts: high-frequency words with the equivalent of more than 500 occurrences, which covered 129 of 51,847 words, and low-frequency words with 500 to 1 occurrences, which covered 51,718 of 51,847 words. We extracted the top ten words in various categories to put more emphasis on the most frequent words used. To begin with, table 2 shows the top ten propositions and conjunctions on the frequency list. In this table, we used many features to extract the top ten, and from these features, we extracted the relative document frequency, which can be expressed as the percentage of documents that contain the item, where the average reduced frequency can be expressed as the variant on a frequency list that discounts multiple occurrences of a word that occur close to each other, and the average logarithmic distance frequency, which is a type of corrected frequency that can be displayed for the results of word lists, n-gram, keywords, and term extraction. The formula for ALDF is

$$ALD = \frac{1}{N} * \sum_{i=1}^{f} di\ log2\ d$$

$$ALDF = 2^{ALD}$$

where N: the size of the corpus, f: frequency of the token, d: distance between the tokens

From this table 2, we found that the first preposition, “with”, can be found in all documents with 100% of the relative document frequency. The average reduced frequency is 19,449, which shows that its frequency from paragraph to paragraph decreased to two-thirds, meaning that the use of this preposition is somewhat clustered in specific regions. Furthermore, we noticed that the difference between the average logarithmic distance frequency value and the frequency value is rather large, which leads to the narrow spread of use of this preposition; the same applies to the following prepositions, where the token is poorly spread throughout the whole corpus.

Table 2. Top 10 prepositions and conjunctions in the corpus: (R) rank; (P) preposition in English equivalent; (F) frequency; (DOCF) document frequency; (RDOCF) relative document frequency; (ARF) average reduced frequency; (ALDF) average logarithmic distance; (CO) conjunction in English equivalent.

R	P	F	DOCF	RDOCF	ARF	ALDF	CO	F	DOCF	RDOCF	ARF	ALDF
1	With/by	30621	1,251	100%	19,449.36	20,168.41	And	64474	1,251	100%	42,727.99	45,030.68
2	In/inside	29056	1,236	98.80%	18,153.89	17,526.83	That / to	17920	1,191	95.20%	10,690.67	10,548.67
3	For /to	27751	1,239	99.04%	17,621.52	18,283.53	What	7173	1,077	86.09%	4,311.20	4,356.43
4	From/since	24040	1,236	98.80%	15,305.64	15,878.54	No/not	4225	838	66.99%	2,141.15	1,904.68
5	On/above	13507	1,216	97.20%	8,330.43	8,519.86	And /so	4032	828	66.19%	2,111.26	1,946.82
6	To/towards/till	8824	1,157	92.49%	5,271.77	5,232.83	After	3172	842	67.31%	1,721.28	1,648.96
7	From /about	5117	1,022	81.69%	3,027.94	3,041.39	Or	2627	683	54.60%	1,173.26	986.28
8	With	4258	946	75.62%	2,349.39	2,281.22	Where	2383	743	59.39%	1,280.91	1,175.49
9	Like /Such as	3214	852	68.11%	1,791.65	1,750.64	But	1638	576	46.04%	866.68	793.94
10	Between/among	2847	817	65.31%	1,521.93	1,481.23	Than/till	1,174	528	42.21%	634.87	599.15

Taking into account that this corpus was collected from ten different newspapers, we can say that each newspaper has its preferences for using a specific preposition set, and each author also has their individual preferences. Conjunctions, on the other hand, show that only the first three commonly used, “and,” “that/to,” and “what”, make up more than 85% of the total; from the fourth onwards it significantly decreases. The difference between the frequency and the ALDF is as large as the proposition distribution, leading to the same results where each author used their preferences for conjunctions, where there are no mutual rules for using them in newspapers. As a result, we can conclude that some newspapers use the token excessively, with a high concentration in small parts, while the rest use it with a good concentration and a reasonable distance between them.

In table 3, we extract the top ten nouns and verbs. It shows that the noun with the highest frequency, “President,” has 937 occurrences, with an insignificant difference from the second noun,

Table 3. Top 10 nouns and verbs in the corpus: (R) rank; (N) noun in English equivalent; (F) frequency; (DOCF) document frequency; (RDOCF) relative document frequency; (ARF) average reduced frequency; (ALDF) average logarithmic distance; (V) verb in English equivalent.

R	N	F	DOCF	RDOCF	ARF	ALDF	V	F	DOCF	RDOCF	ARF	ALDF
1	President	937	401	32.05%	396.56	312.72	To be	2583	678	54.20%	1,310.08	1,207.91
2	The world	934	346	27.66%	405.06	321.71	To say	1472	502	40.13%	632.27	500.75
3	Today	907	468	37.41%	445.55	401.76	To be	1416	580	46.36%	774.71	745.19
4	The president	889	338	27.02%	336.28	241.72	To finish	1240	499	39.89%	567.45	481.70
5	Same	886	437	34.93%	479.86	450.31	To can	945	440	35.17%	472.83	423.12
6	The public	886	461	36.85%	440.66	393.80	To form	807	402	32.13%	417.91	382.33
7	God	833	265	21.18%	264.51	184.74	To finish	682	360	28.78%	330.04	285.49
8	General/public	822	357	28.54%	355.50	299.74	To be	633	370	29.58%	348.45	333.75
9	Yesterday	821	408	32.61%	349.75	191.12	To add	456	293	23.42%	206.06	158.00
10	The work	802	375	29.98%	359.26	307.41	To lift	364	196	15.67%	163.31	137.22

"world." Hereafter, the frequency gradually decreases from one word to another. The table also shows that singular words were more commonly used than dual or plural nouns, plus the determiner "AL" is frequently used with nouns. We can find the same word, "president," in the top ten with both arrangements, with and without the determiner "AL." Furthermore, we can see a significant difference between ALDF and frequency, with political words having a lower relative document frequency than other general words. This means that the majority of the top ten words are political and there is a specific set of political words that are explicitly used in this domain compared to other words. To confirm this conclusion, we can observe that the frequency of the word "today" is lower than the word "president," but its RDOCF is high, and the same goes with the word "same." The other word is "god," which is also used in a narrow domain because its RDOCF is the lowest in the top ten, but its frequency is very high, and its average reduced frequency is weak compared to other ARF values. The results show that many general words exist with an equitable distribution, such as "today," while specific words demonstrate a bias toward specific sets in certain domains. On the other hand, there is no particular selection in verbs, where the most commonly used is the verb "to be", which is the most common in other languages, as well as Arabic. The same applies to the following verbs, with the distribution being relatively even. Nouns are less common than verbs.

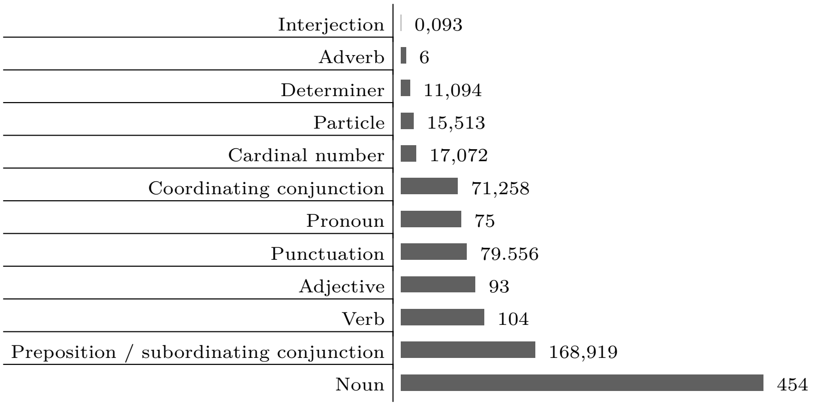

Figure 2. Word frequency distribution for each part-of-speech tag in the proposed corpus.

Figure 2 shows word dispersion by distinguishing labels, with 12 types. This graph shows that the noun labels are the most used in this corpus, with a significant difference from the preposition, conjunction and verb labels. Although the top ten verbs, prepositions, and conjunctions have a much higher frequency than nouns, the set of these labels is limited and has a small value compared to the noun label, which contains a large cluster of words. The following table details the noun label by dividing it into many parts.

Table 4. The part-of-speech tag list of the noun category.

Tag	F	DOCF	RDOCF	ARF	ALDF
Noun/ singular or mass	226,272	1,251	100%	152,361.42	162,739.02
Noun/ singular or mass with the determiner "AL"	133843	1,251	100%	91,529.09	97,849.05
Proper noun/ singular	41,348	1,232	98.48%	19,951.82	17,086.12
Noun/ plural with the determiner "AL"	21,205	1,211	96.80%	12,262.86	11,985.07
Noun/ plural	15723	1,209	96.64%	9,238.58	9,210.09
Proper noun/ singular with the determiner "AL"	9,144	1,092	87.29%	4,626.02	4,151.72
Noun	7542	1,094	87.45%	4,407.16	4,379.56
Proper noun/ plural with the determiner "AL"	20	18	01.44%	8.89	8.53

The noun can be expressed in several ways: singular, plural, and with or without the determiner "AL." We can see that the singular noun with and without a determiner is present in the total corpus, where the difference among the frequency, ARF, and ALDF values is regular but less common when it unites with the determiner. On the other hand, plural words are more commonly used when they are united with a determiner and less commonly used when written as proper nouns. Table 5 details the verb label by dividing it into many parts.

Table 5. The part-of-speech tag list of the verb category.

Tag	F	DOCF	RDOCF	ARF	ALDF
Verb/ Non-3rd Person Singular Present	47,849	1,251	100 %	29,616.04	29,935.07
Verb/ past tense	46,426	1,251	100%	29,487.10	29,824.99
Verb/ past participle	4,283	985	78.74%	2,551.34	2,563.15
Verb/ gerund or present participle	2,051	784	62.67%	1,209.31	1,223.34

The verbs in the proposed corpus were divided into four categories. The first category presents the verb in the non-3rd person singular, with 47,849 frequency and 100% RDOCF. Here, the ARF value makes up more than half of its frequency, indicating it has a widespread presence in the articles. The same applies to verbs in the past tense. Other tenses and forms are less frequently used in the proposed corpus. The lowest category is represented by the base form, which is rarely used in newspaper documents with a 2,051 frequency. However, its RDOCF is 62.67%, meaning that its presence is in most papers but with a rare occurrence. The following table details the adjective label by dividing it into several parts.

Table 6. The part-of-speech tag list of the adjective category.

Tag	F	DOCF	RDOCF	ARF	ALDF
Adjective with the determiner "AL"	54440	1,251	100%	32,605.39	32,195.00
Adjective	33208	1,234	98.64%	19,782.10	19,791.51
Adjective/ comparative	3702	889	71.06%	2,091.11	2,017.13
Adjective/ comparative with the determiner "AL"	1318	570	45.56%	707.91	677.95

There are two types: the simple adjective and the comparative adjective. These two types come with and without the determiner "AL." We find that the simple adjective with the determiner is present in all the documents of this corpus, where its spread of use is good. The same applies to the comparative adjective without the determiner. However, it was not used frequently when accompanied by "AL."

Table 7 details the pronoun label by dividing it into several parts. With 28,846 occurrences of the possessive type and 28,751 occurrences of the personality type, the possessive and personal

Table 7. The part-of-speech tag list of the pronoun category.

Tag	F	DOCF	RDOCF	ARF	ALDF
Possessive pronoun	28846	1,229	98.24%	17,222.51	17,271.79
Personal pronoun	28751	1,222	97.68%	17,173.57	17,128.80
Wh-pronoun	17762	1,210	96.72%	11,205.11	11,594.58

pronouns are used in most articles. Their ALDF and ARF values occupy more than half of the corpus, indicating that the distance among their positions is standard. The wh-pronoun is used less, accounting for about half of the total occurrences.

4.2 N-gram

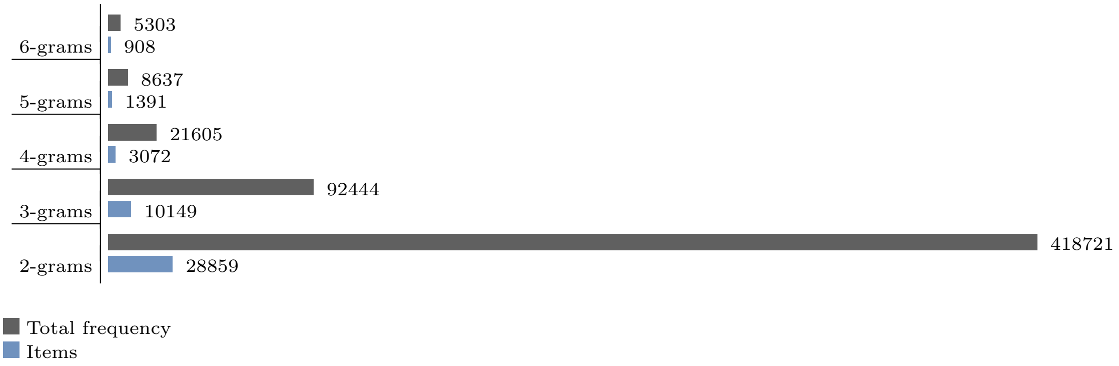

Figure 3. Word Frequency Distribution for N-gram Lengths in the Proposed Corpus.

In this section, we will focus on the N-gram (Almutiri & Nadeem, 2022; Suleiman et al., 2017) and the occurrence probability, which can be very useful for Arabic NLP in auto-completion system predictions, by providing them in 2-gram, 3-gram, 4-gram, 5-gram, and 6-gram statistics. In Figure 3, we outline the n-gram distribution over the corpus, specifying it by two characteristics: the items and their frequency. For the 2-gram, there are 28,859 items with a total frequency of 418,721, while the 3-gram has 10,149 items with a 92,444 frequency; in the 3-gram, the items decreased by two-thirds compared to the 2-gram items, and the frequency significantly reduced from more than 400 thousand to less than 100 thousand.

The same proportion occurred for the 4-gram, 5-gram, and 6-gram, with frequency decreasing in each following contiguous sequence of N items. In the following tables we will outline the top 10 of each gram, and the entire N-gram statistics with features can be found on the following webpage (Bouressace, 2023).

Table 8 illustrates the occurrence of these metrics with their ranks and English equivalents: document frequency, relative document frequency, average reduced frequency, and average logarithmic distance. In the 2-gram set, the first gram represents a combination of three words, “it is a,” whereas, in Arabic it can be just one word with two built-up parts. It is the most frequent 2-gram, with a 2,338 frequency that appeared in more than half of the corpus, plus its ARF and ALDF values took more than half of the frequency, which means it has a regular spread distribution. On the other hand, the other 2-gram frequencies decrease gradually, with the difference from one to another being insignificant. There is no specific term in 2-gram, since all the expressions are general and can be used in any domain.

The most common expression at 3-gram is “and that what,” with 406 occurrences, covering less than a corpus quartile. The contiguous sequence in 3-gram can be just one built-up word

Table 8. Top ten 2- and 3-gram in the corpus with English equivalents.

R	2-gram	F	DOCF	RDOCF	ARF	ALDF	3-gram	F	DOCF	RDOCF	ARF	ALDF
1	It is a	2338	734	58.67%	1,266.02	1,209.11	And that what	406	231	18.47%	204.02	190.54
2	As such	1579	649	51.88%	918.63	917.21	For the	297	192	15.35%	147.89	136.04
3	He is	1452	518	41.41%	693.84	619.46	In addition to	281	200	15.99%	146.44	136.46
4	And he is	1249	529	42.29%	676.54	655.41	As the	246	180	14.39%	139.64	132.75
5	His in	1233	545	43.57%	663.32	581.18	That	220	171	13.67%	125.55	119.93
6	His in	1219	544	43.49%	670.49	653.88	And so	190	147	11.75%	100.92	92.46
7	And in	1107	525	41.97%	622.59	576.42	Because he	179	138	11.03%	101.80	98.89
8	Her and	1066	490	39.17%	558.37	517.66	And others	169	130	10.39%	89.78	87.03
9	That it	1065	505	40.37%	574.22	547.83	It is no	161	125	09.99%	94.16	95.46
10	And by	1019	540	43.17%	598.09	598.22	Until he	128	101	08.07%	61.73	54.98

Table 9. Top ten 4- and 5-gram in the corpus with English equivalents.

R	4-gram	F	DF	RDF	ARF	ALDF	5-gram	F	DF	RDF	ARF	ALDF
1	And others of	82	73	5.84%	47.27	45.98	Peace be upon	55	23	1.84%	12.84	7.16
2	And that's what	69	39	3.12%	25.79	20.22	Be upon him	54	21	1.68%	12.76	7.12
3	For agence france-presse	59	42	3.36%	10.71	3.19	Regional division of the national gendarmerie	25	15	1.20%	6.33	4.93
4	God be upon him and	59	23	1.84%	13.12	7.18	The state department of the judicial police	25	21	1.68%	7.04	5.51
5	May allah bless him	57	24	1.92%	12.98	7.23	Regional national gendarmerie in	25	16	1.28%	6.15	4.87
6	Including	56	43	3.44%	26.06	22.74	In front of the public prosecutor at the court	23	22	1.76%	8.28	6.87
7	Peace be upon him	54	21	1.68%	12.76	7.12	To agence france-presse that	21	18	1.44%	4.15	2.40
8	And that's what made	47	36	2.88%	20.89	17.62	The prophet, may god grant him peace	21	9	0.72%	4.70	3.09
9	It is also	47	40	3.20%	25.97	24.71	The syrian observatory for human rights	19	16	1.28%	5.34	3.28
10	His royal highness prince	43	22	1.76%	6.16	4.45	Thank him and appreciate him	18	12	0.96%	2.22	1.45

or two words. We notice that the greatest frequency is much less than in 2-gram, meaning that the expressions in 3-gram are less frequent. The most frequent are for emphasis, conjunction, and preposition; hence, there is no specific expression in 3-gram. Table 9 illustrates the top ten 4-gram and 5-gram in the proposed corpus. In a 4-gram set, the most frequent expression "and others of" has 82 occurrences, which covers around 5% of the corpus documents. In this contiguous sequence, the terms are specific. According to the top 10, the expressions can be divided into three categories: general fields with five sequences, politics with two sequences, and religion with three sequences. The transition from level to upper causes a change in focus fields, making them more specific and less frequent. These specific expressions cover limited articles, as seen in the ARF values. For the general fields, it gives regular results, which means a steady spread, whereas, in particular areas, it has poor values, which means that the specific expressions are centered on small spaces. Identical concepts existed in the 5-gram, where they became more specific and less frequent. With 1,391 items and 8,637 total occurrences, the most frequent is "peace be upon," with 55 appearing in less than 2% of corpus documents. Here, all the tenth expressions are specific to a specific field; there are expressions related to religion and politics, where the religious are approximately the same in the 4-gram and 5-gram because all the extracted are fragments from the Prophet's Prayer. The top ten 6-grams in the proposed corpus are shown in Table 10.

The 6-gram set contains 908 items with a total frequency of 5,303, where the most frequent one is the expression "peace be upon him," with 51 occurrences appearing in less than 2% of corpus documents. These expressions are similar to those in 3-gram and 4-gram, with vocabulary additions in two fields: politics and religion. The ARF and ALDF values indicate that the 6-gram sequences are localized in one region, meaning these expressions are used by just one or two newspapers. Taking into account that there is no Islamic newspaper, and all of them are political, sports-related, or magazines for health, food, and events, we can say that the Islamic expression "the Prophet's Prayer" was related to a political expression.

Table 10. Top ten 6-gram in the corpus with English equivalents.

R	6-gram	F	DOCF	RDOCF	ARF	ALDF
1	Peace be upon him	51	21	1.68%	12.55	7.08
2	The prophet, may god bless him	21	9	0.72%	4.70	3.09
3	Regional division of the national gendarmerie in	19	14	1.12%	5.39	4.75
4	The covenant, deputy supreme commander, first deputy	17	7	0.56%	2.45	1.99
5	The crown prince, deputy supreme commander, the deputy	17	7	0.56%	2.45	1.99
6	The messenger of god, may god grant him peace	17	8	0.64%	4.13	2.87
7	God, may god grant him peace	16	8	0.64%	4.06	2.86
8	Thank him and appreciate him for	16	10	0.80%	2.09	1.43
9	To the mortuary department of a hospital	15	14	1.12%	5.71	5.24
10	Sheikh tamim bin hamad al thani	14	11	0.88%	4.06	3.68

5 Thematic Vocabulary List

This section presents the vocabulary topics and the most frequently used words in each one. Due to the nature of the corpus, which covered many fields such as world affairs, local affairs, sport, community, health and care, we were able to divide the words into 22 different topics, some of which are rarely used while others are used frequently. The thematic division is derived from previous studies (Abdul Razak, 2011; Buckwalter & Parkinson, 2011), which divided words into specific lists depending on the words' nature. As a result, our division was generated by taking these lists of the words chosen by the researchers of these articles and capturing them in our corpus using the sketch engine tool. The primary aim of extracting this thematic vocabulary list is to gain a deeper understanding of the content within the newspapers and to analyze the central themes within each topic. For each topic, our goal is to identify the most frequently used words. Through this analysis, we aim to determine the extent to which the words used within the newspapers accurately represent and cover their respective topics. The following graph (Figure 4) illustrates the most common topics.

In Figure 4, we notice that the most frequent topic is politics and law, while the least frequent topic is animals. Word lists have been compiled for each topic, ordered by frequency. These lists include multi-context words; words that share the same letters but shift in meaning between topics. In Arabic, a single word often has several meanings, leading to potential misinterpretation when moving between topics. For instance, a word with high frequency in one topic may not exclusively pertain to that topic, as other contexts might influence its frequency. The top 15 words for each topic are presented in the following graphs, selected in order to analyze the most important words within each topic and to specify the true frequency of the words. This number is ideal as many topics have a maximum of 15 words, ensuring consistency (Bouressace, 2023).

Figure 5.1 displays the frequency distribution of words in the animal topic, highlighting the top 15 words. It reveals that only 14 words existed with a low frequency of 177 occurrences. This topic is infrequently mentioned in world affairs, sports, or community news. Additionally, due to varying usage, several words were not included in these statistics, including names of specific persons or places. For instance, the word “falcon,” with four occurrences in this corpus, serves as a proper name rather than referring to animals. In Figure 5.2, the most frequent word is “shirt,” with 42 occurrences, while the least frequent word is “laundry,” making a single appearance. This topic is also rarely addressed in the newspapers, and many words have meanings that diverge from their expected context. For example, the word ’headband,’ with 87 occurrences, is used to describe a “gang.” Similarly, the word “patterned” can mean “plan.” Moving to the color topic (Fig. 5.3), we have identified 20 specific words with a total of 895 occurrences. The most frequent word is ’white,’ with 179 appearances. However, due to the broad nature of this topic for this type of newspaper or magazine, some words have multiple meanings. For instance, “brown” can refer to a color or “child,” and “green” has another interpretation.

Figure 6.1 shows the distribution of words by frequency in the Emotion topic, which includes 59 different words with a total frequency of 744. The most frequent word is “angry,” with 68 occurrences. This topic is not used so frequently in this corpus, in which frequent words, such as

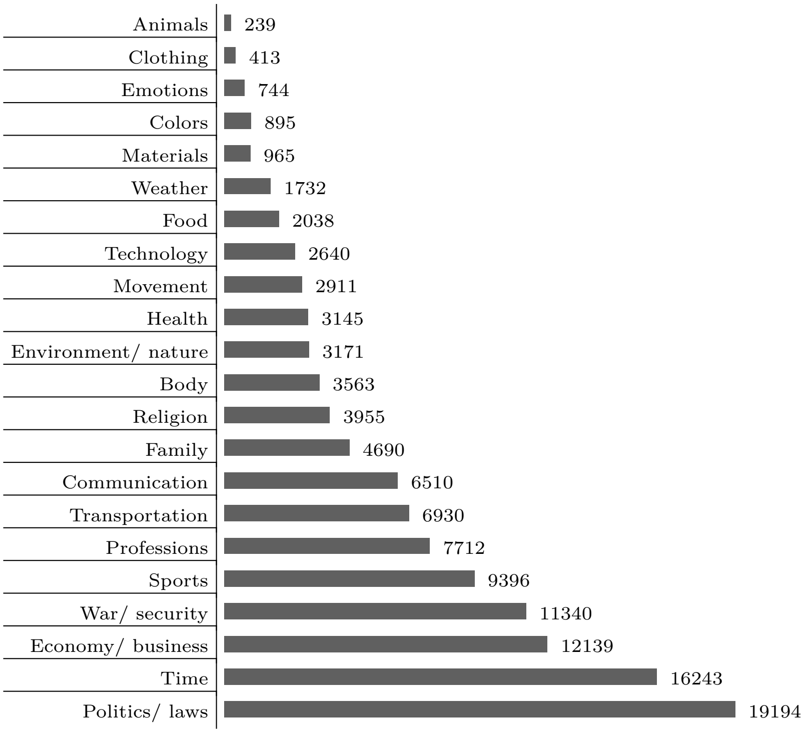

Figure 4. Dispersion of the most common topic frequency.

"happy" and "safe" (used as proper nouns) and "be accustomed to" (used as "back to") are not related to the topic. The communication words in Figure 6.2 are frequently used in many places, with 6,510 occurrences in 61 items; the most frequent word is "to say," with 1472 occurrences. Most of the existing words in this topic are contextually accurate, and the first word is prominently utilized compared to the subsequent words. This pattern suggests that world affairs articles consider this word fundamental for conveying information.

Figure 6.3 depicts the family topic, which has 4,690 occurrences in 68 items, with the most frequent word being "family" with 333 frequencies. This topic is frequently used in corpus articles, but many words listed in this topic have various meanings.

Figure 7.1 illustrates the distribution of words by frequency in the war and security topic, encompassing 162 words with a total frequency of 11,340. The most frequently occurring term is "security," with 809 occurrences. Within the proposed corpus, this topic ranks as the fourth most frequently discussed. It holds a direct connection to world affairs newspapers, and various expressions employing these words are prevalent. Similarly, the politics and law topic comprises 152 items totaling 19,194 occurrences, making it the most extensively covered topic. The word "president" leads with 1,826 occurrences, making it the most frequent term in the entire corpus.

Figure 7.3 graphically presents words related to the religion topic. This category contains 91 different words, totaling 3,955 occurrences. The word "God" stands out as the most frequently used with 833 instances. The majority of words in this topic are used appropriately in the religious context, although there are a few with alternate meanings.

Figure 8.1 shows the distribution of words by frequency in the environment and nature topic, which includes 59 different words with a total frequency of 3,171. The most frequent word, with 262 occurrences, is "environment." This topic is frequently used in articles, where many words have different meanings, such as "canal," which can also mean "channel". The weather vocabulary word list is depicted in Figure 8.2, consisting of a total of 1,732 frequencies across 43 items. The most frequently occurring word in this category is "season" with 486 instances. While this topic

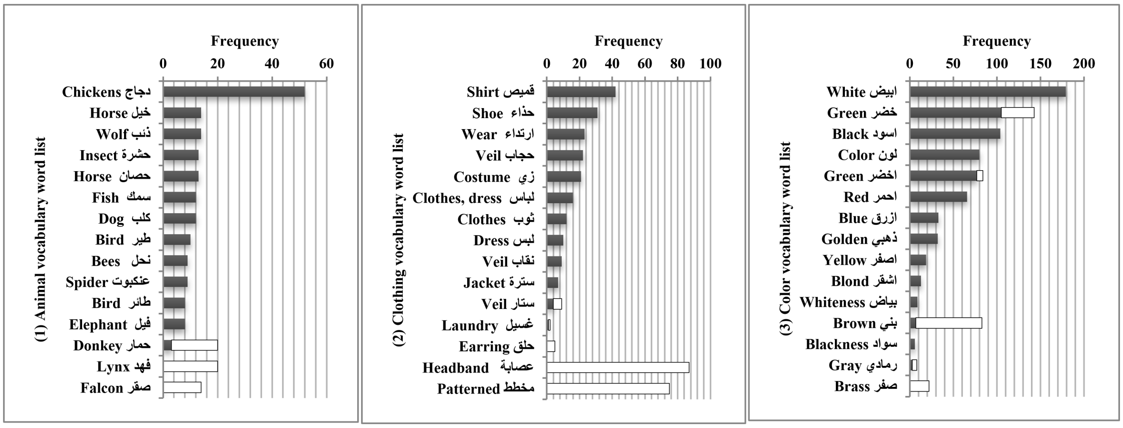

Figure 5. Dispersion of perfect and imperfect animal, clothing, and color vocabulary word lists.

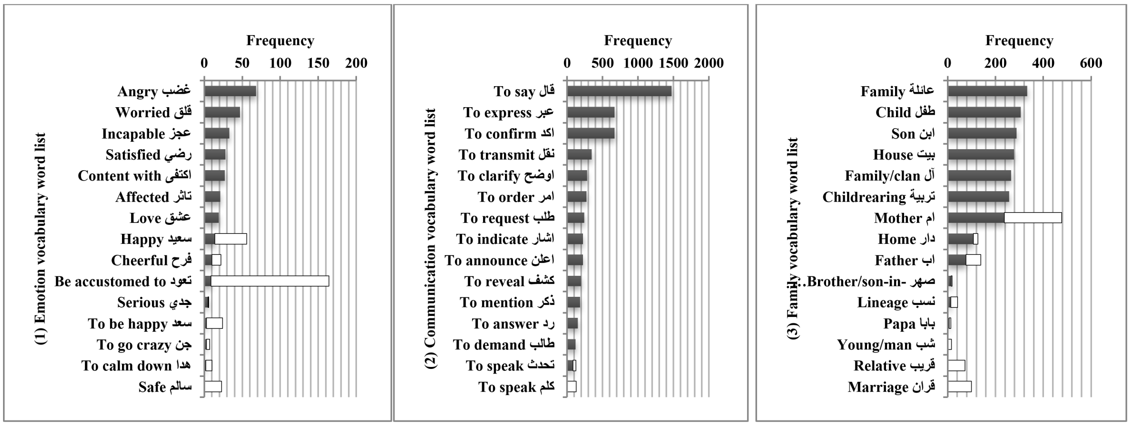

Figure 6. Dispersion of perfect and imperfect emotion, communication, and family vocabulary word lists.

is not extensively used across most newspaper domains, a significant portion of the words are appropriately employed within the context of weather-related articles. Figure 8.3 represents the transportation topic, which encompasses 6,930 occurrences and 92 items. The most commonly used word in this category is “to cross,” with a frequency of 670. This vocabulary list exhibits a large cluster of words that have been utilized in the articles and stands out for having a substantial number of words compared to many other topics. Furthermore, it contains words with multi-context usage, encompassing both accurate contextual meanings and alternative interpretations.

Figure 9.1 shows the distribution of words by frequency in the health topic, which includes 50 different terms with a total frequency of 3,145. The most common word, with 339 occurrences, is “infection.” This topic is not frequently used in the articles, in which health words have a weak impact, meaning that only a few words have a high frequency while the majority have a small frequency number. Figure 9.2 represents the food vocabulary list, which has 2,038 occurrences of 70 items. Many words in this topic have meanings other than those used for this topic, such as “sugar” (which can mean “close”) and “salad”. The body words in Figure 9.3 have 3,563 occurrences using 48 items, with the most common, “foot”, having 445 frequencies in accurate meaning and 201 frequencies in false meaning, indicating that most body words have multiple meanings.

Figure 10.1 illustrates the distribution of words by frequency in the material topic, which includes 39 different words with a total frequency of 965. The most common word, with 108 occurrences, is “petroleum.” This topic is limited to some words with low frequency. The same

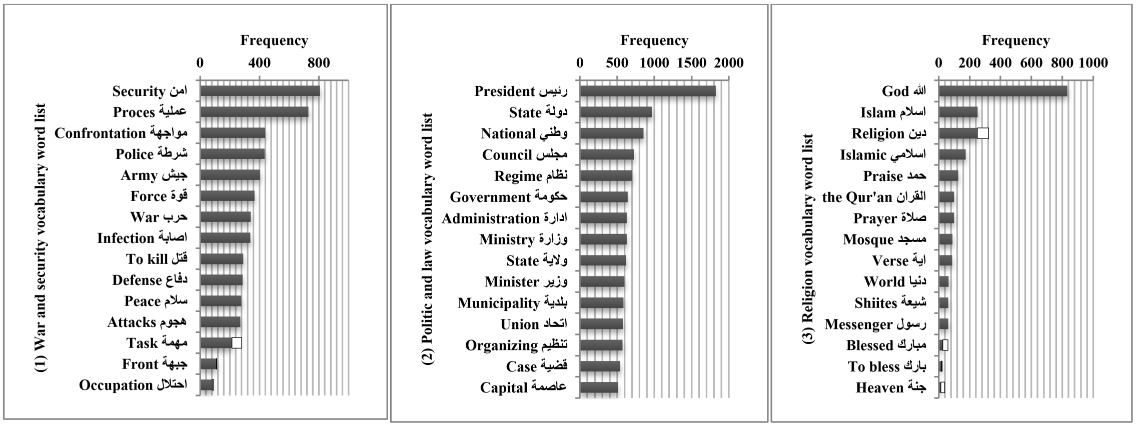

Figure 7. Dispersion of perfect and imperfect war and security, politic/ law, and, religion vocabulary word lists.

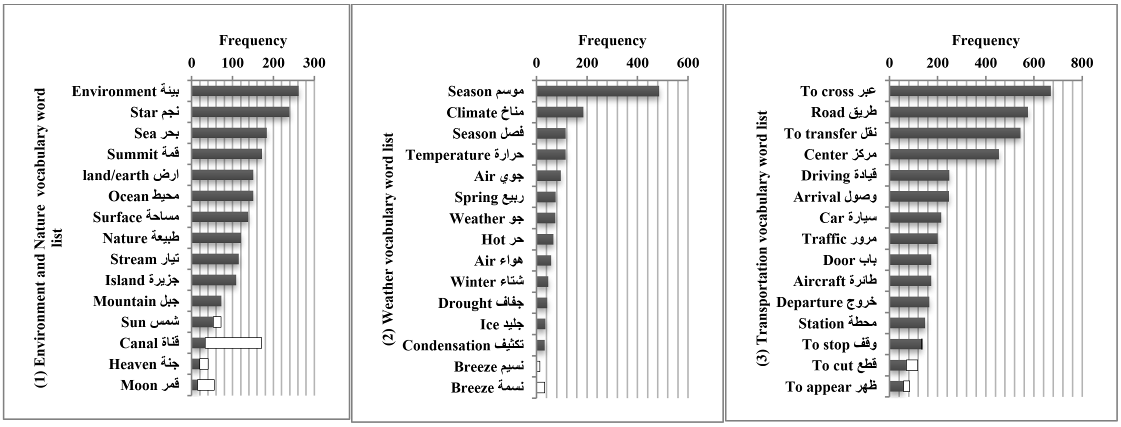

Figure 8. Dispersion of perfect and imperfect environment/ nature, weather, and transportation vocabulary word lists.

goes for the technology vocabulary word list, with a total of 2,640 frequencies using 49 items, where the most frequent word in this topic is “program”, with 353 occurrences.

Figure 11.1 illustrates the distribution of words by frequency in the movement topic. We have identified 83 words with a total of 2,911 occurrences, where “to advance” is the most frequent word, appearing 302 times. This topic is infrequently discussed in different newspaper sections. Additionally, some terms were not included in these statistics due to their dissimilar usage. The sports vocabulary word list in the corpus articles is well-represented, with 9,396 occurrences using just 62 words. “team” stands out as the most common word with 1,059 occurrences. The words in this topic have a significant number of occurrences, with the top 15 words in the graph having more than 6,600 occurrences. In Figure 11.3, the time vocabulary word list contains 16,243 occurrences across 126 words, with “day” being the most common word at 1,527 occurrences. This topic is the second most frequent in this corpus, primarily due to the abundance of words used across various fields, including sports and world affairs. Many words with multiple meanings contribute to the frequency values, such as “period,” which can mean “yes,” and “noon,” which can mean “back” or “to appear”.

Figure 12.1 shows the distribution of words by frequency in the profession topic. We were able to extract 124 words with a total of 7,712 occurrences, with “president” being the most frequent

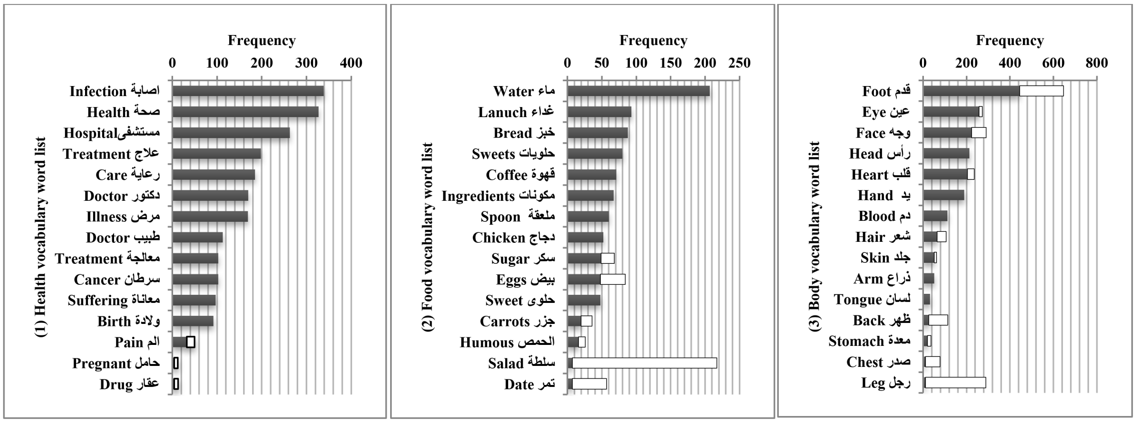

Figure 9. Dispersion of perfect and imperfect health, food, and body vocabulary word lists.

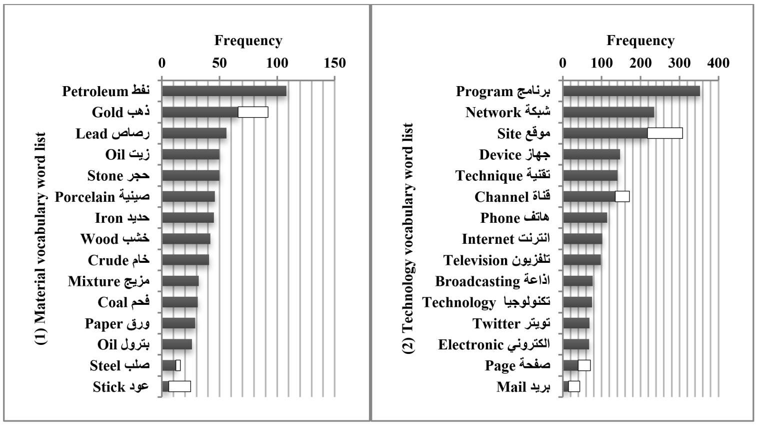

Figure 10. Dispersion of perfect and imperfect material and technology vocabulary word lists.

word with 1,826 occurrences and considered the most frequent word in this corpus (the word with and without the determiner “AL”). This word is so frequent compared to the following terms due to the political meaning that can be attached to it and because it is used in various fields in the articles. The second graph (Figure 12.2) includes the economic and business vocabulary word list. We could extract 137 words with a total frequency of 12,139, where the most frequent word is “system,” with 701 occurrences. This topic is the third most frequent in this corpus, with convergent values from the first to the 14th word, then significantly decreasing to small values.

6 Conclusions

In this paper, we presented a new frequency dictionary of Arabic extracted from a printed Arabic text database (PATD) by studying 1,251 XML files, 2,344 articles, and 1,000,000 words from ten Arabic newspapers collected from their official online websites, giving a total of 51,847 frequently used words, 19,926 sentences, and 22 topics. This dictionary contains a wealth of information for each word in the most frequently used word lists, such as the English equivalents of each sample, usage statistics, and usage distribution across several important Arabic articles. The frequency dictionary is arranged by a frequency index and contains all the useful Arabic vocabulary to make it easier for researchers and learners. The dictionary is freely available at Bouressace (2023)

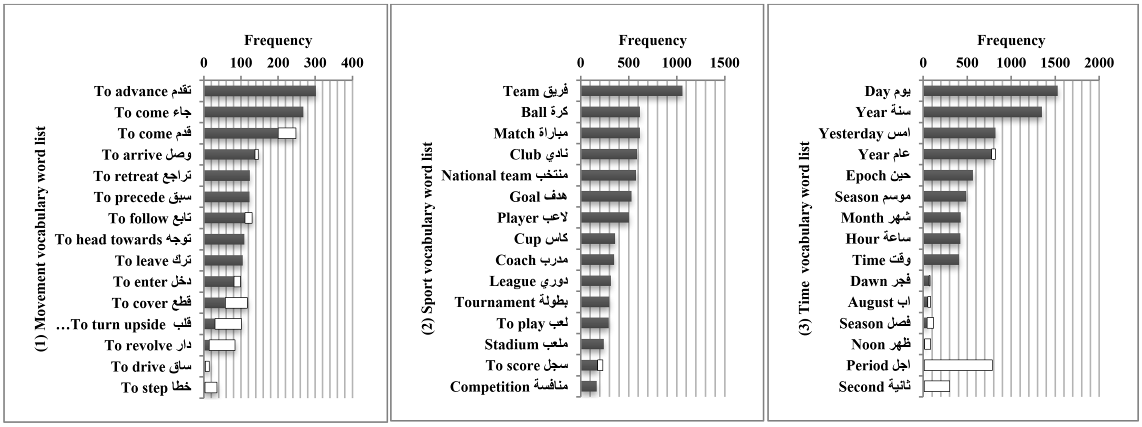

Figure 11. Dispersion of perfect and imperfect movement, sport, and time vocabulary word lists.

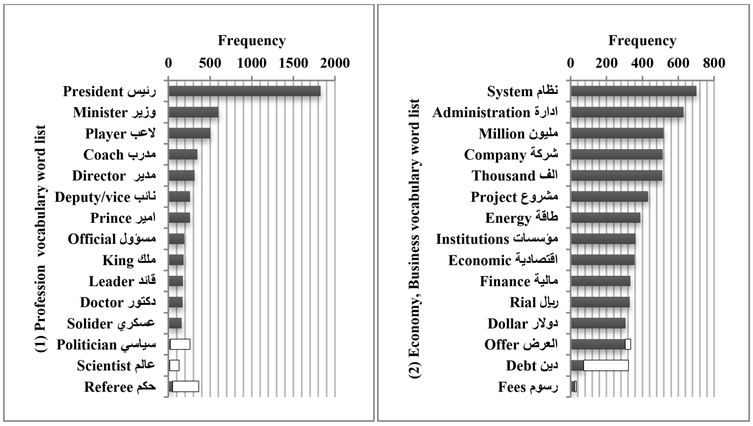

Figure 12. Dispersion of perfect and imperfect profession, economy, and business vocabulary word lists.

for research purposes, with the hope that it will assist the Arabic language processing research community in such areas as indexing language and retrieval language.

In this study, we systematically analyzed data across various phases, considering different aspects. We commenced with an examination of frequency distributions and explored how these distributions influenced the entire corpus. Our corpus consisted of a recurring set of words found in articles covering diverse topics, including politics, sports, and social issues. Additionally, we investigated the frequencies of their respective lexical categories. This analysis allowed us to pinpoint essential words within each category, greatly contributing to various linguistic applications, such as language indexing and retrieval. Furthermore, we delved into N-grams, which revealed that larger N-grams led to more precise yet less frequent selections. This observation led us to infer that phrases composed of the same words are often reserved for special occasions, events, or expressions of particular significance within Arabic society. We further scrutinized these words using the thematic vocabulary list to assess their frequency in specific contexts. This list aided in determining whether a word is context-specific or versatile, indicating multiple meanings across various fields, including sports, politics, body and health, and more. Some words were specific to a single category, while others had the capacity to transcend boundaries, assuming different meanings in various contexts. This analysis assisted us in distinguishing between words with consistent frequencies within a single context and those with multiple contexts because a high frequency of a word does not necessarily signify its reference to something special. Our investigation revealed that a substantial portion of words fell within the one-context category, denoting a singular meaning within their respective fields. However, we also identified a smaller subset of words with multi-contextual flexibility, capable of assuming diverse meanings in various contexts.

References

Abdelali, A. (2003). Localization in modern standard Arabic. *Journal of the American Society for Information Science and Technology*, *55*(1), 23–28. https://doi.org/10.1002/asi.10340

Abdelali, A., Cowie, J., & Soliman, H. S. (2005). *Building a modern standard Arabic corpus: Paper presented at the Computational Modeling of Lexical Acquisition Workshop, Croatia, 25th to 28th of July*. https://www.researchgate.net/publication/228958341_Building_a_modern_standard_Arabic_corpus

Abdul Razak, Z. R. (2011). *Modern media Arabic: A study of word frequency in world affairs and sports sections in Arabic newspapers* [Doctoral dissertation, University of Birmingham]. https://etheses.bham.ac.uk/id/eprint/2882/

Abuleil, S., & Evans, M. (2002). Extracting an Arabic lexicon from Arabic newspaper text. *Journal of Computer and the Humanities*, *36*(2), 191–221. https://doi.org/10.1023/A:1014368121689

Adham, M. A. A., al-Angelo, A. M., Agresti, A. N. D., & Finlay, B. (2009). *Statistical methods for the social sciences* (4th ed.). Pearson Education.

Alderson, J. C. (2007). Judging the frequency of English words. *Applied Linguistics*, *28*(3), 383–409. https://doi.org/10.1093/applin/amm024

Alhaj, Y. A., Wickramaarachchi, W. U., Hussain, A., Al-Qaness, M. A. A., & Abdelaal, H. M. (2018). Efficient feature representation based on the effect of words frequency for Arabic documents classification. In *Proceedings of the 2nd International Conference on Telecommunications and Communication Engineering (ICTCE 2018)* (pp. 397–401). Association for Computing Machinery. https://doi.org/10.1145/3291842.3291900

Almutiri, T., & Nadeem, F. (2022). Markov models applications in natural language processing: A survey. *International Journal of Information Technology and Computer Science (IJITCS)*, *14*(2), 1–16. https://doi.org/10.5815/ijitcs.2022.02.01

Alshammari, R. (2018). Arabic text categorization using machine learning approaches. *International Journal of Advanced Computer Science and Applications*, *9*(3). 226–230. Retrieved May 25, 2019, from https://doi.org/10.14569/IJACSA.2018.090332

Al-Sulaiti, L., & Atwell, E. (2006). The design of a corpus of contemporary Arabic. *International Journal of Corpus Linguistics*, *11*(2), 135–171. https://doi.org/10.1075/ijcl.11.2.02als

Ayadi, R., Maraoui, M., & Zrigui, M. (2016). A survey of Arabic text representation and classification methods. *Research in Computer Science*, *117*, 51–62.

Bouressace, H. (2023). *A frequency dictionary of printed Arabic text*. http://www.inf.u-szeged.hu/patd/fdpatd/

Bouressace, H., & Csirik, J. (2019). Printed Arabic text database for automatic recognition systems. In *Proceedings of the 2019 5th International Conference on Computer and Technology Applications (ICCTA '19)* (pp. 107–111). Association for Computing Machinery. https://doi.org/10.1145/3323933.3324082

Buckwalter, T., & Parkinson, D. (2011). *A frequency dictionary of Arabic – core vocabulary for learners: Edition bilingue anglais-arabe*. Routledge.

Dornyei, Z. (2007). *Research methods in applied linguistics: quantitative, qualitative, and mixed methodologies*. Oxford University Press.

Duwairi, R., Al-Refai, M. N., & Khasawneh, N. (2009). Feature reduction techniques for Arabic text categorization. *Journal of the American Society for Information Science and Technology*, *60*(11), 2347–2352. https://doi.org/10.1002/asi.21173

El Kourdi, M., Bensaid, A., & Rachidi, T. (2004). Automatic Arabic document categorization based on the naïve Bayes algorithm. In *Proceedings of the Workshop on Computational Approaches to Arabic Script-based Languages* (pp. 51–58). COLING. https://doi.org/10.3115/1621804.1621819

Goweder, A., & De Roeck, A. N. (2001, July 6). Assessment of a significant Arabic corpus. In *Proceedings of the Arabic NLP Workshop at ACL/EACL 2001: ARABIC Language Processing: Status and Prospects*.

https://www.researchgate.net/publication/233967788_Assessment_of_a_significant_Arabic_corpus

Kilgarriff, A., Baisa, V., Bušta, J., Jakubíček, M., Kovář, V., Michelfeit, J., Rychlý, P., & Suchomel, V. (2014). The Sketch Engine: Ten years on. *Lexicography*, *1*(1), 17–36. https://doi.org/10.1007/s40607-014-0009-9

Kilgarriff, A., Rychlý, P., Smrž, P., & Tugwell, D. (2004). The Sketch Engine. In *Proceedings of the 11th EURALEX International Congress* (pp. 105–116). Universite de Bretagne-Sud.

Masrai, A., & Milton, J. (2016). How different is Arabic from other languages? The relationship between word frequency and lexical coverage. *Journal of Applied Linguistics and Language Research*, *3*(1), 15–35.

Mesleh, A. M. A. (2007). Chi Square Feature Extraction Based SVMs Arabic Language Text Categorization System. *Journal of Computer Science*, *3*(6), 430–435. https://doi.org/10.3844/jcssp.2007.430.435

Suleiman, D., Awajan, A., & Al Etaiwi, W. (2017). The use of hidden Markov model in natural ARABIC language processing: A survey. *Procedia Computer Science*, *113*, 240–247. https://doi.org/10.1016/j.procs.2017.08.363

Syiam, M., Fayed, Z., & Habib, M. (2006). An intelligent system for Arabic text categorization. *International Journal of Intelligent Computing and Information Sciences*, *6*(1), 1–19.

Uysal, A. K., & Gunal, S. (2012). A novel probabilistic feature selection method for text classification. *Knowledge-Based Systems*, *36*, 226–235. https://doi.org/10.1016/j.knosys.2012.06.005

The publication was financed at the author's expense.

The author declares that she has no competing interests.

Publisher: Institute of Slavic Studies, Polish Academy of Sciences

Publishing history: Received 2023-04-02; Accepted 2023-10-30; Published 2023-12-31.